Review of “Neuron Shapley: Discovering the Responsible Neurons”

Authors: Amirata Ghorbani & James Zou @ Stanford

Presenter: Ning Zhang @ CUHK-SZ

Mar 16, 2021

Interpretation of Neural Networks

1. **Feature importance:** contribution of each input feature
   - Integrated Gradient [1], DeepLIFT [2], LIME [3], etc.
2. **Sample importance:** contribution of each training example
   - Data valuation based on Shapley values [4, 5], etc.
3. **Element importance:** contribution of each neuron/filter
   - Neuron conductance [6], extension of feature importance methods [7], etc.

Contributions of the Paper

1. **Conceptual:** develop Neuron Shapley framework to quantify the contribution of each neuron/filter.
2. **Algorithmic:** introduce multi-arm bandit based algorithm to efficiently estimate Shapley values.
3. **Empirical:** apply the results to facilitate model interpretation and repair regarding accuracy, fairness, robustness, etc.

Preliminaries

Notation

- **Model:** network $\mathcal{M}$ with L layers each with $n_{l\in\{1,\cdots,L\}}$ neurons/filters. $N = \{m_i\}_{i=1}^{n}$ includes the $n = \sum_{l=1}^{L} n_L$ individual elements of $\mathcal{M}$.
- **Performance:** $V(N)$ assigns score to trained network, e.g. accuracy, loss.
- **Contribution:** m_i's contribution towards $V(N)$ is denoted as $\phi_i(V, N)$, satisfying $\sum_{i=1}^{n} \phi_i(V, N) = V(N)$ with $\phi_i = \phi_i(V, N)$ for simplicity.

Desirable Properties for ϕ_i

- **Zero Contribution:** $\phi_i = 0$ if $V(S \cup \{i\}) = V(S)$ for all $S \subset N - \{i\}$.
- **Symmetry:** $\phi_i = \phi_j$ if $V(S \cup \{i\}) = V(S \cup \{j\})$ for all $S \subset N - \{i, j\}$.
- **Additivity:** when $V = V_1 + V_2$, $\phi_i(V, N) = \phi_i(V_1, N) + \phi_i(V_2, N)$.

Shapley Value in Neuron Evaluation

The authors propose the **Neuron Shapley** ϕ_i as

$$\phi_i = \frac{1}{|N|} \sum_{S \subset N - \{i\}} [V(S \cup \{i\}) - V(S)] / \binom{|N|-1}{|S|}.$$

- Computation of $V(S)$: for neurons, output of $N \backslash S$ can be replaced by zeros. For filters, output of $N \backslash S$ are replaced by mean outputs for validation images.
- Take into account the interactions between neurons.
- Uniquely satisfy the three aforementioned desirable properties.

Shapley Value in Game Theory [8]

Cooperative game with a set N of n players and reward function $v : 2^n \to \mathbb{R}$.

- $v(S)$: reward the members of S can obtain by cooperation with $S \subset N$.
- ϕ_i distributes the group reward among players in an equitable way.
- Equitable way means the aforementioned desirable properties are satisfied.

Estimation of Shapley Value

Monte-Carlo Estimation

- The Shapley value of the i-th element [4] can be written as

$$\phi_i = \mathbb{E}_{\pi\sim\Pi}[V(S_\pi^i \cup \{i\}) - V(S_\pi^i)],$$

π: permutation of $\{1, \cdots, n\}$, Π: uniform distribution over $n!$ permutations, S_π^i: set of elements that appear after i in given permutation π.

- Approximating ϕ_i can be transformed to estimating the mean of certain r.v..

Early Truncation Technique

- When $S_\pi^i \cup \{i\}$ is small, $V(S_\pi^i \cup \{i\})$ can degenerate to negligible due to loss of connections, i.e. $V(S_\pi^i \cup \{i\}) - V(S_\pi^i) \approx 0$ in this case.
- Alternatively, choose a performance threshold v_T, and if $V(S_\pi^i \cup \{i\}) < v_T$, set $V(S_\pi^i \cup \{i\}) - V(S_\pi^i) = 0$.

Algorithm: Truncated Multi Armed Bandit Shapley (TMAB-Shapley)

Input: Elements $N = \{1, \cdots, n\}$, metric $V(\cdot)$, tolerance ϵ, number of important elements k, performance threshold v_T

Initialization: $\{\phi_i\}_{i=1}^n = 0$, $\{\sigma_i\}_{i=1}^n = 0$, $\mathcal{U} = N$, $t = 0$

while $\mathcal{U} \neq \emptyset$ **do**
- $t = t + 1$, random permutation π^t of $\{1, \cdots, n\}$
- **for** $j \in \mathcal{U}$ **do**
  - **if** $V(S_{\pi^t}^j \cup \{j\}) < v_T$ **then**
    - $\phi_j^t = 0$
  - **else**
    - $\phi_j^t = V(S_{\pi^t}^j \cup \{j\}) - V(S_{\pi^t}^j)$
  - **end**
  - ϕ_j = Moving Average(ϕ_j^t), σ_j = Moving Variance(ϕ_j^t)
  - $(\phi_j^{lb}, \phi_j^{ub})$ = Confidence Bound(ϕ_j, σ_j)
- **end**
- $\mathcal{U} = \{i : \phi_i^{lb} + \epsilon <$ top k largest $\phi_i < \phi_i^{ub} - \epsilon\}$

end

Output: Shapley values $\{\phi_i\}_{i=1}^n$

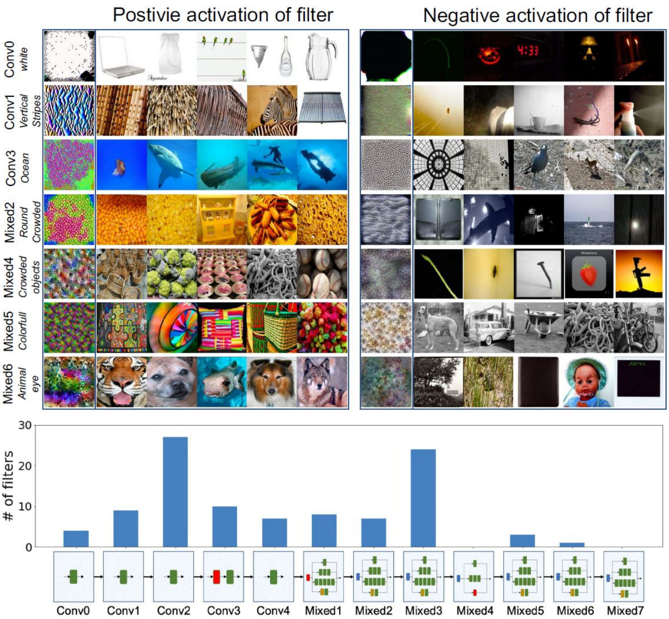
Application I: Identifying Critical Filters for Overall Accuracy

Implementation details

- **Model:** Inception-v3 network [9] with 17216 filters preceding the logit layer trained on ImageNet [10]. Its validation set is divided into two parts (25000 images each) to serve as validation (for zero out) and test sets.
- **Method:** TMAB-Shapley with $V(\cdot)$ set to the overall accuracy of the network on a randomly sampled batch of images and $k = 100$.

Interpretation Results

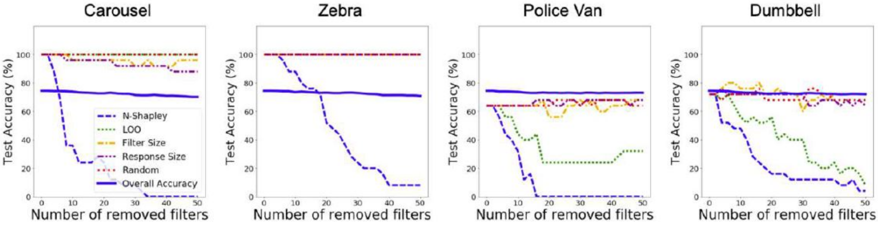
- Neuron Shapley values are very sparse, most of them are close to zero.
- Remove top 10 filters, overall test accuracy drops from 74% to 38%; Remove top 20 filters it drops to 8%. Random removal does not change the accuracy.
- Visualize the filter with the highest Shapley value in 7 of the layers applying:
  - Deep Dream [11]: image optimized by gradient ascend to highly activate filter.
  - Five images in the validation set that highly activate the filter.



Postivie activation of filter
Negative activation of filter
Conv0 white
Conv1 Vertical Stripes
Conv3 Ocean
Mixed2 Round Crowded
Mixed4 Crowded objects
Mixed5 Colorfull
Mixed6 Animal eye
# of filters
0
10
20
30
Conv0
Conv1
Conv2
Conv3
Conv4
Mixed1
Mixed2
Mixed3
Mixed4
Mixed5
Mixed6
Mixed7

Application II: Identifying Critical Filters for Chosen Class

- **Aim:** Identify filters that highly contribute to the chosen class but are not crucial for the overall performance.
- **Method:** TMAB-Shapley with $V(\cdot)$ set to class recall. Exclude the top 20% filters that contribute mostly to the overall accuracy from above experiment.
- **Results:** Removing top 40 filters leads to a dramatic decline in the network's ability to detect that class, but the overall performance remains intact.

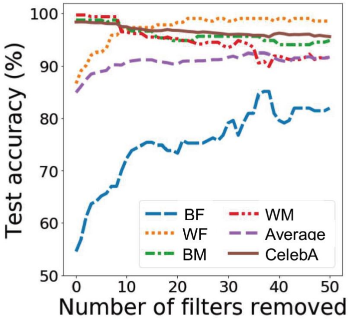


Visualization of Filters with the Highest Shapley Values

- Deep Dream: Image optimized to achieve the positive activation of filters.
- Top five images in the training set that highly activate the filters.

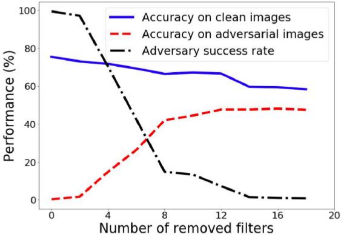
Application III: Discovering Unfair Filters

- **Motivation:** Gender detection models have certain biases towards minorities, e.g., they are less accurate on black female faces.
- **Model:** SqueezeNet [12] with 2976 filters trained on the celebA [13] dataset.
- **Method:** $V(\cdot)$ is set to accuracy on the PPB dataset [14].
- **Results:** Zeroing out filters with most negative Shapley values greatly increases the accuracy on black female (BF) faces from 54.7% to 81.9%, and also improves that for white females (WF).
- The average accuracy on PPB increased from 84.9% to 91.7%.
- The performance on male faces and CelebA data only drops a little after modification.



Application IV: Improving Adversarial Robustness

- **Goal:** Identify filters vulnerable to adversaries in the Inception-v3 model.
- **Adversaries:** Use iterative PGD attack [15] as an adversary to perturb each validation image so it's misclassified as a randomly chosen class.
- **Method:** $V(\cdot) =$ Adversary's success rate $-$ Accuracy on clean images. Former is the rate of fooling the network predicting randomly chosen labels.
- **Results:** Zeroing out the filters with the top 16 Shapley values, the adversary's attack success rate drops from nearly 100% to 0.1%, while the model's performance on clean images drops from 74% to 67%.
- The network is robust to the original adversary, but still vulnerable to new adversaries designed for the new model.
- For black-box adversaries created by different architectures, their attack success rate drops by 37% on average.



Main References

[1] Sundararajan, M., Taly, A. and Yan, Q., 2017, July. Axiomatic attribution for deep networks. In *International Conference on Machine Learning* (pp. 3319-3328). PMLR.

[2] Shrikumar, A., Greenside, P. and Kundaje, A., 2017, July. Learning important features through propagating activation differences. In *International Conference on Machine Learning* (pp. 3145-3153). PMLR.

[3] Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

[4] Ghorbani, A. and Zou, J., 2019, May. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning* (pp. 2242-2251). PMLR.

[5] Jia, R., Dao, D., Wang, B., Hubis, F.A., Hynes, N., Gürel, N.M., Li, B., Zhang, C., Song, D. and Spanos, C.J., 2019, April. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 1167-1176). PMLR.

[6] Dhamdhere, K., Sundararajan, M. and Yan, Q., 2018. How important is a neuron?. *arXiv preprint arXiv:1805.12233*.

[7] Datta, A., Sen, S. and Zick, Y., 2016, May. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)* (pp. 598-617). IEEE.

[8] Shapley, L.S., 1953. A value for n-person games. *Contributions to the Theory of Games*, 2(28), pp.307-317.

Main References

[9] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).

[10] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A.C., 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), pp.211-252.

[11] Olah, C., Mordvintsev, A. and Schubert, L., 2017. Feature visualization. *Distill*, 2(11), p.e7.

[12] Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J. and Keutzer, K., 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 MB model size. *arXiv preprint arXiv:1602.07360*.

[13] Liu, Z., Luo, P., Wang, X. and Tang, X., 2018. Large-scale celebfaces attributes (celeba) dataset. *Retrieved* August, 15(2018), p.11.

[14] Buolamwini, J. and Gebru, T., 2018, January. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

[15] Kurakin, A., Goodfellow, I. and Bengio, S., 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.